Evaluating Administrative Records to Inform Measurement Error Properties of National Survey of College Graduates Estimates: An Analysis of the NSCG-LEHD Earnings Ratio

by

Michaela Dillon
U.S. Census Bureau

CARRA Working Paper 18-13 September, 2021

This paper was written in 2018 under the former Center for Administrative Records Research and Applications program, but not published until 2021. CARRA has since been combined with the Center for Economic Studies, and the CARRA working paper series has been discontinued.

To obtain information about the series, see www.census.gov/ces or contact Christopher Goetz, Editor, Discussion Papers, U.S. Census Bureau, Center for Economic Studies 5K038E, 4600 Silver Hill Road, Washington, DC 20233, CES.Working.Papers@census.gov.

Abstract

Using administrative records in survey operations can potentially improve data accuracy and survey operations. In this study, we link administrative data on earnings from the Longitudinal Employer-Household Dynamics (LEHD) dataset to the National Survey of College Graduates (NSCG) to understand the alignment of this administrative records information with respondent collected data. Around 50 percent of linked individuals report earnings in the NSCG that are within ten percent of their LEHD earnings. Large disagreement between linked values appears to be prevalent among high-earning individuals and demonstrates some association with characteristics of low labor market attachment such as part time status and retirement age.

Keyword*:* administrative records, college graduates, earnings

JEL Classification: C83, E2, C8

[*] This paper is released to inform interested parties of research and to encourage discussion. The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau. All results have been reviewed to ensure that no confidential information is disclosed, DRB approval number CBDRB-FY18-433.

EXECUTIVE SUMMARY

Using administrative records in survey operations can potentially improve data accuracy and reduce respondent burden. In this study, we link administrative data on earnings from the Longitudinal Employer-Household Dynamics (LEHD) dataset to the National Survey of College Graduates (NSCG) to understand the alignment of this administrative records (AR) information with respondent collected data.

The LEHD program populates its database with state unemployment insurance (UI) and Quarterly Census of Employment and Wages (QCEW) data linked to other administrative records and Census Bureau data. This information allows the tracking of both aggregate and individual level employment, earnings, and job flows over time. Established in the 1970s, the NSCG is a biennial survey that collects data on the college-educated population of the United States, highlighting the connection between educational attainment and subsequent labor market outcomes.

This research evaluates conceptual alignment, linkage, and agreement of annual earnings data between data sources and their stability across certain respondent characteristics. Upon linking the LEHD to the NSCG by Protected Identification Key (PIK), the data shows generous coverage of the NSCG sample, over 90 percent. Among linked cases, agreement is assessed by examining percent differences in linked values and by the ratio of NSCG to LEHD earnings. Overall, 11.01 percent of the linked LEHD data is within one percent of the NSCG value—35.34 percent and 53.02 percent of the linked data is within five percent and ten percent of the NSCG value, respectively. Disagreement between linked values appears to be prevalent among high-earning individuals, and demonstrates some association with characteristics of low labor market attachment such as part time status and retirement age.

Due to its ability to inform the direction and magnitude of disagreement between linked earnings values, the earnings ratio represents measurement error in this analysis. The distribution of the earnings ratio suggests the NSCG has a slight tendency to under-report the AR value. Additionally, regression analysis conducted on this ratio yields statistically significant effects from earnings, sex, age, educational attainment, job tenure, and self-employment status. In particular, the earnings differential decreases at an increasing rate as earnings rise—evidence of NSCG under-reporting among richer respondents. Alternatively, employer changes, self-employment, and higher educational attainment increase the differential. These findings are robust within several subsample estimations.

Some limitations affect the potential of this AR for survey data replacement and/or supplementation. Specifically, since the LEHD relies on unemployment insurance records the LEHD may not fully cover marginal workers such as the very young or old, retirees, students, or self-employed. Difficulty also arises when trying to link survey respondents that simply choose not to work. Additionally, federal employees are underrepresented in the LEHD and available earnings data includes different elements slightly inconsistent with wage data used for non-federal employees.

Despite these limitations, the LEHD is a highly useful data source for earnings information among formally employed individuals of prime working age. Future research comparing these results to those using federal tax data for robustness can strengthen confidence in these findings. An additional set of recommended analyses would investigate the sensitivity of the earnings distribution to replacing survey data with AR values can help determine the extent to which under-coverage issues distort the analysis.

Table of Contents

List of Figures and Tables

I. INTRODUCTION

The National Survey of College Graduates (NSCG) is a longitudinal survey of the college-educated population living in the United States. Sponsoered by the National Center for Science and Engineering Statistics (NCSES) within the National Science Foundation (NSF), the survey informs two congressionally-mandated reports, *Women, Minorities, and Persons with Disabilities in Science and Engineering and Science* and *Engineering Indicators*, on the composition and productivity of the nation's STEM workforce. Thus, NCSES, with the Census Bureau serving as the data collection contractor, administers the NSCG to collect information on the human capital investment decisions and labor market outcomes of highly-educated workers. Over time, the survey tracks respondents' demographic characteristics, educational attainment, workplace training, job satisfaction, professional mobility, and income.

NCSES is interested in the use of administrative data sources to enhance and supplement NSCG information. Administrative records have the potential to address many goals, including: informing on measurement error, supplementing respondent-collected data, and reducing data collection and processing costs. To that end, NCSES has requested that the Demographic Research Area in the Center for Economic Studies (CES) of the Census Bureau (formerly the Center for Administrative Records Research and Applications) to evaluate the NSCG for the use of administrative records to supplement and/or replace items on its questionnaire.

One promising source of administrative data with potential value to NSCG is the Census Bureau's Longitudinal Employer-Household Dynamic Program (LEHD). The LEHD consists of several files structured at both the individual and firm/establishment level of analysis. The LEHD program gathers and organizes employment, earnings, and firm-level information over time from various sources including state agencies, business surveys and censuses, and federal tax forms. These data are used to improve imputation methods and to generate synthetic data for testing purposes with respect to employment information. Generally, historic data is available for all states from year 2000 to the most recent data release. Overall, the LEHD accounts for over 1.5 billion jobs, held by 262 million people, across 21 million firms.[1]

In this study, CES will assess the coverage, agreement, and quality of available LEHD data to respondent-provided information in the 2010 NSCG, specifically focusing on measurement error in annual earnings.[2] This memo addresses several objectives in the assessment of LEHD data for potential NSCG item supplementation or replacement. First, this work will identify which NSCG employment and earnings items could be enhanced by the information available within LEHD. Second, it will measure the extent to which measurement error exists within the data as determined by (dis)agreement between linked LEHD values and the distribution of responses

[1] Full documentation of the LEHD program is available here: https://www2.census.gov/ces/wp/2014/CES-WP-14-26.pdf.

[2] This paper is the result of continuing research evaluating the use of LEHD data to enhance the NSCG. Prior findings on LEHD in relation to several NSCG employment history survey items including primary salary are found in Dillon, M. (20XX) "Evaluating Administrative Records to Inform Measurement Error Properties of National Survey of College Graduates Estimates: Employment History and Firm Characteristics."

between the two datasets. Finally, we will determine if data quality varies by key employment and demographic characteristics.

II. LITERATURE REVIEW

This research investigates the ability of administrative records to replace and/or enhance survey data on annual earnings. Specifically, the analysis includes identifying appropriate administrative data, linking it to survey responses, and evaluating the extent of measurement error of earnings data between the two sources. The benefits to using AR with survey data are numerous. Künn (2015) emphasizes the reliability of AR to researchers as a routinely collected and authoritative data source. AR potentially lower data collection costs, circumvent the need for multiple datasets, and increase research opportunities via enhanced datasets. Additionally, the use of AR has been studied as a method to increase sampling efficiency for certain subpopulations, validate survey data, supplement survey data for difficult to obtain information, and improve forecasting ability of program costs (Bowie and Kasprzyk, 1987). An especially important benefit of AR is its ability to address various types of error within survey data. For example, the use of AR is one strategy to minimize item nonresponse, the instance when a respondent does not answer certain questions on the survey.

Sensitive topics such as earnings and income exhibit relatively larger rates of item nonresponse across household surveys (Meyer et al., 2015). Reasons for this behavior include concerns for privacy, cognitive difficulty/lack of understanding of survey item, risk of program eligibility, off-the-books earnings, and stigma of certain responses (Kunn, 2015; Bollinger et al., 2015). Because of respondent selection into nonresponse status or even intentional misreporting, bias within the data is not necessarily random across respondent characteristics. In fact, nonresponse rates take on a U-shape across the income distribution where respondents at either end tend to omit income (Lillard et al., 1986; Bollinger et al., 2015; Korinek, 2005). High nonresponse rates are also associated with certain occupations where the calculation of net income for tax purposes may be ambiguous, such as farmers and private household workers at low-income levels, and lawyers at upper income levels. In addition, educational attainment and work experience increases income nonresponse (Lillard et al., 1986).

Another form of error addressed by AR is measurement error, the difference between the survey value and its true value. Comparison of linked survey and AR values sharing conceptual alignment sheds light on the extent to which measurement error is evident in survey values. For instance, large disagreement in linked values signals increased measurement error. Like item nonresponse, measurement error varies by respondent characteristics. Several studies find evidence of underreporting of earnings as respondent income rises. Likewise, earnings are overreported in survey data among low-income respondents (Bee, 2013; Roemer, 2002; Brummet et al., 2017). In CPS data, part-year and part- time workers tend to underreport earnings relative to AR, while workers with multiple jobs overestimate earnings (Roemer, 2002). Kreuter et al. (2014) observe measurement error among “hard to recruit” respondents already reluctant to participate in the survey. Specific to repeated surveys, Bollinger and David (2001) find measurement error is higher for respondents that do not respond to all waves of a panel survey. Additionally, measurement error within longitudinal data is positively autocorrelated, suggesting persistence in this error over time (Bound and Krueger, 1991).

III. DATA

3.1 National Survey of College Graduates (NSCG):

The NSCG is a biennial survey sponsored by the National Center for Science and Engineering Statistics (NCSES) within the National Science Foundation, administered by the Census Bureau, and sampled from the American Community Survey (ACS). It implements a rotating panel design in which respondents answer questions about their employment status, earnings, and education up to four times over a period of about six years. One of the unique features of the NSCG is its collection of data on more subjective information such as motivating factors for the individual's human capital investments, change in career or employment status. Additionally, the information collected in this survey informs two congressionally mandated reports on the U.S. STEM labor force: *Women, Minorities, and Persons with Disabilities in Science and Engineering*, and *Science and Engineering Indicators*. Survey respondents are college graduates, living in the U.S., up to age 75.

This study uses 2010 NSCG restricted access data. This particular year of the survey was the first data release after switching to its current sample frame, the ACS. To maintain the continuity of the rotating panel design, 46,828 new observations from the 2009 ACS were added to the sample already including 30,360 return respondents sampled from the 2001-2008 panels of the National Survey of Recent College Graduates and the 2003 NSCG for a total of 77,188 observations.

3.2 Longitudinal Employer-Household Dynamics (LEHD) Program:

The LEHD dataset is a collection of standardized data files sourced from administrative records on local employment, businesses, and earnings. Specifically, the data comes from state-level unemployment insurance (UI) records, the Quarterly Census of Employment and Wages (QCEW), as well as other administrative records sources available within the Census Bureau. The files link together via unique person and firm/establishment identifiers, the protected identification key (PIK) and state employer identification number (SEIN), respectively. Linkage allows the tracking of both aggregate and individual level employment, earnings, and job flows over time. The availability of this information varies by state depending on the data use agreement between the state data owners and the Census Bureau. The LEHD has nearly universal coverage of all states from the year 2000 forward.

This study uses the Employment History Files (EHF) from the 2011 LEHD snapshot that provide quarterly earnings information for every job held by a worker as far back as year 1990 in some states. Therefore, each observation is a person-employer-year record of earnings. The universe numbers over three trillion observations. LEHD generates a record within the EHF only if the employer paid positive, nonzero wages during any quarter that year. Therefore, there are no missing values for annual earnings, the sum of quarterly earnings from an employer in a particular year, in the EHF.

3.3 Limitations

The LEHD files are an important resource that provides researchers a detailed description of worker employment histories. These histories, however, only include jobs and earnings reported to government agencies. Therefore, there is a coverage issue with LEHD data among certain occupations with informal payment practices. The LEHD underreports employment activity for those Abraham et al. (2013) characterize as marginal workers such as the very young or old, retirees, students, or self-employed. Difficulty also arises when trying to link survey respondents that simply choose not to work, since only the employed are present in the LEHD.

Additionally, federal employees are underrepresented in the LEHD EHF files. The LEHD program addresses this limitation by harmonizing, to the extent possible, separate records from the Office of Personnel Management (OPM) which are interleaved into the EHF files by PIK in this analysis.. These federal employee earnings data include different elements which are slightly inconsistent with wage data from state UI records, therefore introducing some wrinkles in conceptual alignment of earnings between the data sources. Also, a number of federal agencies, mostly national defense and justice, do not contribute data to this framework, citing security issues.

Lastly, states may join or opt out from the LEHD program at any time. Therefore, depending on the status of the individual agreement states have with the Census Bureau, LEHD's coverage by state may fluctuate over time.

IV. RESEARCH QUESTIONS

The research questions are as follows:

1. To what extent are the concepts measured by the NSCG earnings question *aligned* with the administrative record information?
2. How often do NSCG records *link* to viable administrative record data that can be used to replace or supplement survey responses?
3. How often do data from the administrative records source *agree* with the responses from NSCG respondents by major subpopulation characteristics?

V. METHODOLOGY

The research questions of the previous section correspond to three analytical objectives of this research. That is, to assess linkage, conceptual alignment, and agreement of annual earnings information between the NSCG and LEHD. This section presents supplemental information on the analysis used to produce the data in the results section.

5.1 Conceptual Alignment

For research question #1, evaluation of conceptual alignment involves verifying the data collected within both data sources are as similar as possible. In this study, that includes a comparison of the LEHD and NSCG definitions of annual earnings across three data sources: NSCG, state UI records, and OPM data for federal employees. Conceptual alignment also includes manipulating LEHD data when necessary in order to provide the closest approximation

to requested NSCG information. This process is described in the following data management section. Finally, the analysis includes comparison of the annual earnings distributions for the survey and administrative records data via kernel density estimates. This visualization highlights irregularities between the data sources as well as any outliers that will help guide research efforts to any areas of concern.

5.2 Linkage:

For research question #2, in order to link the NSCG to the LEHD, both datasets require assignment of the unique identifier for individuals, the Protected Identification Key (PIK). PIKs allow linkage of information for a particular person across various Census surveys and administrative records. PIKs assignment to datasets occurs via the Personal Identification Validation System (PVS), a probabilistic matching algorithm used to anonymize incoming data at the Census Bureau. This process uses personally identifiable information (PII) from the survey such as name, age, and address to search reference files containing all known transactions for an SSN. Once matching information is found in the reference files with a certain threshold of confidence, the unique PIK value replaces PII found on the survey data file. See Wagner and Layne (2014) for a detailed description of the PVS process. The linkage rate, based on unique PIKs, represents the proportion of the PIKed NSCG sample found in the LEHD database. The analysis also includes the calculation of the linkage rate across various respondent demographics.

5.3 Agreement:

For question #3, the analysis includes findings on 1) agreement in response value, and 2) the behavior of measurement error across respondent characteristics. First, plots of the average percent difference in linked values over the LEHD earnings distribution provides high-level overview of (dis)agreement and the relative relationship of linked values at different levels of income. Second, a table showing the distribution of percent differences between linked earnings data provides insight on the frequency of percent differences between linked values. Third, to highlight the relative relationship between linked values, a second table presents the distribution of the NSCG-LEHD earnings ratio. Due to its ability to inform on the direction and magnitude of disagreement between linked earnings values, the earnings ratio represents measurement error in this analysis. In both tables, the analysis replicates these distributions across several socio-demographic characteristics including gender, age, race/ethnicity, citizenship, educational attainment and (non)STEM background. Disaggregating the overall distributions allows observation of potential determinants of measurement error as well as identification of outliers within the data.

Finally, regression analysis on the log NSCG-LEHD earnings ratio as a function of several respondent socioeconomic and employment characteristics tests the statistical significance of any observed variances in measurement error. The empirical model is below. Tables in the results contain a list of all regressors.

$$log\left(\frac{NSCG\ annual\ earnings}{LEHD\ annual\ earnings}\right) = \alpha + \beta \cdot LEHD\ earnings\ quintile + \gamma \cdot demographic + \delta \cdot employment + \varphi \cdot education + \varepsilon$$

The results contain benchmark estimates for this model and for several subsamples. In particular, the discussion compares benchmark results to those of non-marginal workers (i.e. workers with high labor market attachment such as full time employment, prime working age, not self-employed), workers aged under 30 years, 30-49 years, 50-64 years, and 65+ years, full time, part time, STEM and non-STEM employees.

VI. DATA MANAGEMENT

The 2010 NSCG contains 77,000 observations. After undergoing PVS processing to assign PIKs to this dataset, the PIK rate is 98.35% (76,000 unique PIKs). Survey data typically has high PIK rates (90-93%). Failure to receive a PIK often occurs when SSN is unknown and/or disconnected from government programs and records. Additionally, non-PIKed persons are likely to be a non-U.S. citizen, be unemployed, not have health insurance, live in poverty, be under 35 years of age, be a minority, or have less than a high school education (NORC, 2011). Since individuals of high socioeconomic status, such as college graduates, exhibit fewer of these characteristics, the higher than average PIK rate for the 2010 NSCG is justified.

Next, the assigned PIKs were merged onto the 2010 NSCG response file via the survey unique identifier, REFID. After identifying unique PIKs from the NSCG file, it was linked to the LEHD Employment History File (EHF), which is a person-job-level file. Therefore, each observation within the EHF represents the PIK-SEIN-SEINUNIT-YEAR (person-firm-establishment-year) combination, wherein quarterly earnings information is available for each job held by an individual.[3] Because an individual may hold more than one job in a year, this merge is a one-to-many match, resulting in a linked dataset of approximately 1.2 million observations.

The NSCG survey item requests a single value per person for annual earnings. Deriving an equivalent annual earnings value from the multiyear job-level linked dataset requires the following steps:[4]

1. From the linked dataset, drop records for years other than 2009. This deletion results in a dataset including all jobs for individuals employed in 2009.
2. For each PIK, create a running total of annual earnings for each job so that the last observation for each PIK is the aggregated 2009 annual earnings across all jobs for each person.
3. Collapse the dataset to unique PIKs by retaining the final observation for each PIK, which reports the derived annual earnings value. Delete all other repeated PIK observations.

After removing NSCG records with missing or duplicate PIKs and linking to EHF records from 2009, the resulting dataset includes nearly 52,000 unique PIK observations.[5]

[3] State Employer Identification Number (SEIN). SEINUNIT is an establishment-level identifier.
[4] A visualization of this exercise is available in the appendix.
[5] Of the 77,188 observations in the 2010 NSCG, 1.5% were not assigned a PIK, 0.2% were duplicate PIKs, and 31.0% did not link to 2009 EHF data (were not employed according to LEHD).

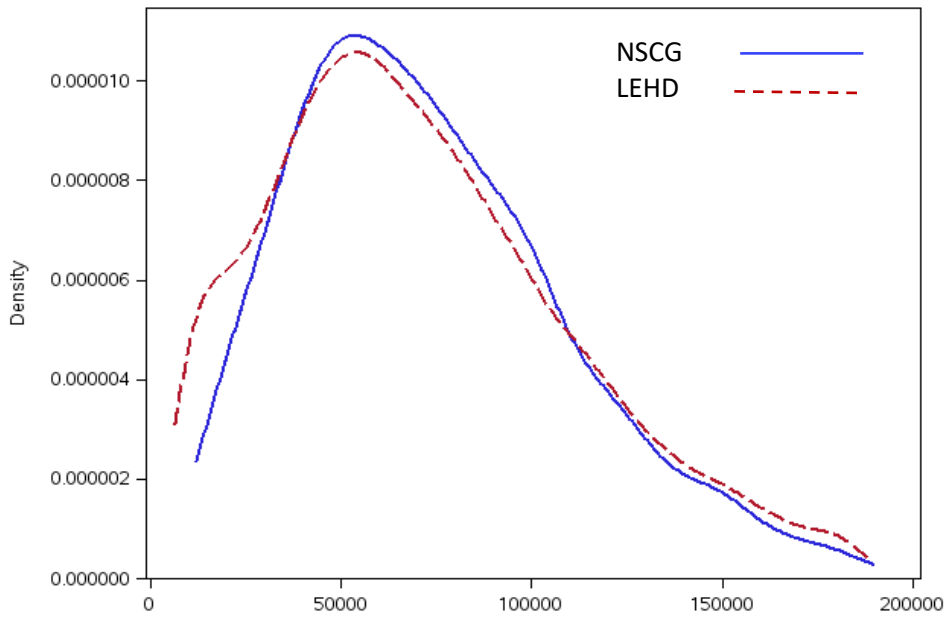
VII. RESULTS[6]

7.1 Conceptual Alignment

Item A38 on the 2010 NSCG questionnaire asks respondents, "*Counting all jobs held in 2009, what was your total earned income for 2009, before deductions? Include all wages, salaries, bonuses, overtime, commissions, consulting fees, net income from businesses, summertime teaching or research, or other work associated with scholarships.*" The response is our variable of interest for 2009 annual earnings, EARN.[7]

LEHD EHF wage data includes UI-covered earnings reported by an employer (Vilhuber and McKinney, 2014). According to the BLS Handbook of Methods, UI wages include *gross wages and salaries, bonuses, stock options, tips and other gratuities, and the value of meals and lodging* (BLS, 1997). Separate files produced by the U.S. Office of Personnel Management (OPM) contain wage data for federal employees. Specifically, the data reflects quarterly earnings derived from a "total pay" variable including *basic pay, locality adjustment, supervisory differential, retention allowance, and cost of living allowance*. Some adjustments outlined in Vilhuber and McKinney (2014) are applied to the OPM data to create closer alignment with the UI data in the EHF files. These consistent UI and OPM earnings measures include all wage and salary income reported to the government and provide the closest approximation to the NSCG definition of annual earnings among formally employed individuals.

Figure 1: Kernel Density Estimates of LEHD and NSCG 2009 Annual Earnings



Source: 2010 NSCG linked to 2011 LEHD EHF files by PIK. N=51,000.

[6] Please note all results in tables are rounded or suppressed where necessary for disclosure avoidance to protect respondent privacy.

[7] The item nonresponse rate for this variable is twelve percent (NCSES, 2010). About eight percent of the values for EARN are imputed in the linked sample used for analysis.

After deriving a consistent annual earnings value from the LEHD data as described in the data management section, Figure 1 shows the overall distributions of LEHD and NSCG earnings aligns rather well across most earnings values. The main difference in the distributions occurs towards the left tail where LEHD more frequently reports low earnings values. This result foreshadows possible over-reporting of survey earnings values among low-income respondents—a finding documented in existing research discussed in the literature review. This pattern persists among different subsamples of the dataset.

7.2 Linkage

Linkage to the LEHD EHF files resulted in a linked dataset of approximately 1.2 million person-job-level observations. Collapsing the merged dataset to unique PIKs shows 93.95 percent of the PIKed NSCG sample as employed at some point in time covered within the LEHD.

As shown in Table 1, the linkage rate varies over some demographic characteristics. These results provide insight into the degree to which AR provides coverage of certain groups of people. Looking at the first two columns, variation in the linkage rate occurs over age groups ranging from 86.50 percent for workers age 65-75 to 96.30 percent for workers in their twenties. Other respondent characteristics that display noticeable difference in linkage rates are ethnicity and citizenship. Hispanic (87.08%) and non-U.S. citizen (89.35%) respondents link less frequently to LEHD than non-Hispanics (94.69%) and those with U.S. citizenship (94.32%).

Table 1: NSCG-LEHD Linkage Rates Across Respondent Characteristics

	Full Dataset		Selected Subset	
	Count	Linkage Rate	Count	Linkage Rate
Overall	**76,000**	**93.95**	**52,500**	**95.60**
Male	42,500	94.17	28,000	96.42
Female	33,000	93.68	24,500	94.65
21-29	8,700	96.30	6,600	96.58
30-39	17,500	96.88	14,500	96.91
40-49	17,500	95.29	13,500	95.91
50-64	24,000	92.60	18,000	93.98
65-75	8,200	86.50		-
Asian	12,000	94.61	8,700	95.44
Black	7,200	95.67	5,100	96.92
Multiple race	1,900	94.20	1,400	95.69
AIAN	400	95.68	300	96.72
NHPI	350	95.59	250	96.27
White	54,000	93.55	37,000	95.43
Hispanic	7,300	87.08	5,400	88.25
Non-Hispanic	68,500	94.69	47,500	96.44

Table 1: NSCG-LEHD Linkage Rates Across Respondent Characteristics

	Full Dataset		Selected Subset	
	Count	Linkage Rate	Count	Linkage Rate
U.S. citizen	70,500	94.32	48,000	96.10
Not a U.S. citizen	5,600	89.35	4,500	90.17

Source: 2010 NSCG and 2011 LEHD EHF files.
Note: In light of LEHD under-coverage of workers with low labor market attachment in the private sector, we show the linkage rate among a selected subset excluding respondents not of prime working age, employed part-time due to retiree or student status, a federal employee, or self-employed. Results rounded or suppressed (D) where necessary for disclosure avoidance.

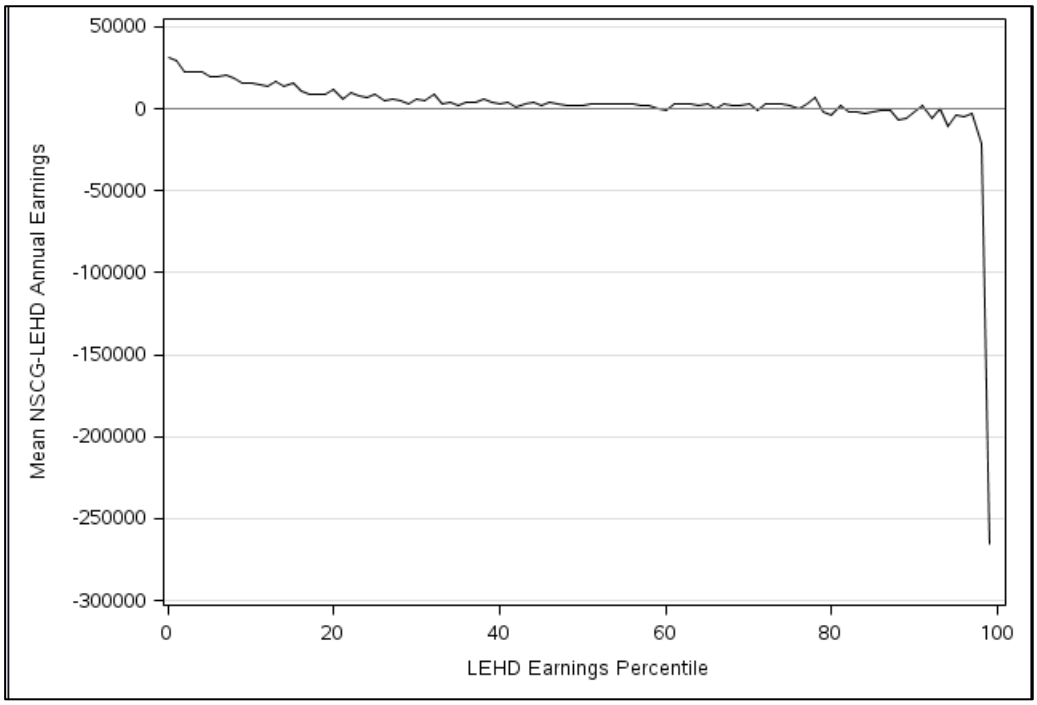
Abraham et al. (2013) acknowledge under-coverage of certain workers by the LEHD exists. The last two columns of Table 1 recalculate the linkage rate within a subsample excluding "marginal" workers as a robustness check. Specifically, the reduced sample excludes workers age 65 or older, retirees, students, self-employed, and federal employees. Consequently, the overall linkage rate slightly improves to 95.60 percent. Likewise, the linkage rates across all the listed characteristics rise and maintain the relative relationships across demographic categories found in the full sample. Note the particularly low linkage rates among older workers and non-citizens. Most workers begin to consider retirement in their sixties prompting an exit from the labor force, and non-citizens sometimes face barriers to entering the labor force due to visa restrictions. The following results uncover similar underperformance in agreement among certain marginal workers with low labor market attachment.

7.3 Agreement

7.3.1 Overview

Figure 2 shows the mean difference (NSCG minus LEHD) between linked earnings values within percentiles of the LEHD earnings distribution, and confirms the notion from Figure 1 that NSCG values exceed their corresponding LEHD value at the low end of the distribution. The positive difference persists to around the 60th percentile. Among the wealthiest respondents in the top five percent, LEHD earnings values far exceed survey values. The general trend of the difference approaching negative values as income rises agrees with findings of previous research observing increased nonresponse and underreporting of income among high-income individuals. In fact, this visualization very closely resembles the distribution of linked Consumer Expenditure Survey and W-2 earnings data studied in Brummet et al. (2017). Part of the large differences seen at the top of the income distribution is related to the NSCG survey instrument itself. The NSCG survey instrument caps earnings values just shy of $10 million, whereas LEHD data does not impose such constraints. Nonetheless, underreporting is prevalent even for high earners with earnings below this topcode, so the survey instrument is not the sole driver of differences at the top of the distribution.

Figure 2: Average Difference between LEHD and NSCG Annual Earnings by Percentile of the LEHD Earnings Distribution



Source: 2010 NSCG linked to 2011 LEHD EHF files by PIK. N=100.

7.3.2 Percent Difference in Linked Values by Respondent Characteristics

Table 2 shows the distribution of several ranges of percent difference in order to highlight the frequency of linkages to similar values within the data as well as linkages to outlier values. Overall, 11.01 percent of the linked LEHD data is within one percent of the NSCG value—35.34 percent and 53.02 percent of the linked data is within five percent and ten percent of the NSCG value, respectively. The modal range of agreement between linked values is 2 to 5 percent, representing 24.33 percent of the data. This pattern persists across most respondent characteristics.

Table 2: Agreement Rates of Linked Annual Earnings Values across Employment and Demographic Characteristics

		Distribution of Agreement					
	N	within 1%	2-5%	6-10%	11-20%	21-50%	over 50%
Overall	51,000	11.01	24.33	17.68	16.75	16.47	13.76
Sex							
Male	29,000	11.12	24.64	17.66	16.71	16.18	13.67
Female	21,500	10.86	23.91	17.62	16.89	16.83	13.89
Age							
Age: Less than 30	5,800	9.56	22.60	16.56	15.99	17.08	18.20
Age: 30-49	25,500	10.90	25.14	18.93	17.53	15.64	11.86
Age: 50-64	17,000	11.58	24.82	16.75	16.33	16.82	13.70
Age: 65+	2,700	11.65	17.35	13.29	14.41	20.66	22.64
Race/Ethnicity							
White	31,500	11.23	25.30	17.91	16.66	15.82	13.08
Black	4,800	10.23	23.65	17.68	16.04	17.61	14.79
Asian	8,300	11.24	22.35	17.18	17.26	16.83	15.14
AIAN	200	7.58	20.85	14.69	18.48	22.27	16.11
NHPI	250	9.33	22.67	18.67	14.67	20.44	14.22
Multiple race	1,000	10.27	25.22	15.05	16.75	16.15	16.55
Hispanic (of any race)	4,700	10.33	22.05	17.25	17.67	18.55	14.15
Citizenship							
U.S. citizen	47,000	10.99	24.44	17.77	16.78	16.36	13.66
Non-citizen	3,600	11.27	22.92	16.05	16.93	17.75	15.09
Educational Attainment							
Undergraduate degree	27,500	10.98	24.08	18.11	17.47	16.11	13.24
Graduate degree	23,500	11.05	24.62	17.09	15.99	16.87	14.38
Hours Worked[1]							
Full time	41,500	11.38	26.40	18.79	17.28	15.23	10.93
Part time	5,500	8.58	14.42	12.71	15.26	22.57	26.45
STEM Occupation							
STEM job (narrow)	23,000	11.94	26.98	19.14	17.18	14.97	9.80
Non-STEM job	28,000	10.26	22.18	16.43	16.48	17.67	16.99
STEM job (broad)	29,000	11.62	26.43	18.80	17.08	15.26	10.82
Non-STEM job	22,000	10.21	21.58	16.12	16.42	18.04	17.64
Earnings Distribution Position							
LEHD earnings quintile 1	10,000	6.45	9.50	7.88	11.40	22.41	42.37
Quintile 2	10,000	11.11	26.20	20.01	17.57	17.47	7.63

Table 2: Agreement Rates of Linked Annual Earnings Values across Employment and Demographic Characteristics

		Distribution of Agreement					
	N	within 1%	2-5%	6-10%	11-20%	21-50%	over 50%
Quintile 3	10,000	12.98	30.52	21.12	18.11	13.06	4.21
Quintile 4	10,000	14.14	31.06	21.57	17.52	11.52	4.18
Quintile 5	10,000	10.37	24.38	17.63	19.35	17.83	10.43

Source: 2010 NSCG and 2011 LEHD EHF files.
1:Total excludes respondents not working during reference week of Oct. 1, 2008.
Results rounded or suppressed (D) where necessary for disclosure avoidance.

Deviations from the average distribution occur among the youngest and oldest age groups, part time workers, and low-income individuals. Each of these characteristics relate to relatively low (or diminishing in the case of older workers) labor market attachment. Interestingly, the previous discussion on LEHD linkage limitations and alignment highlighted the potential of lower linkage rates for these groups. Each of these groups are more likely to link to LEHD values more than 50 percent different from their survey responses. In fact, that is the modal rate of agreement for these groups, with the exception of very young workers. Low earning individuals up to the 20th percentile display the greatest tendency (42.37%) to link within this range. Groups of considerably smaller sample size, such as the AIAN and NHPI populations, also display noticeable differences, redistributing agreement from within 1 percent into the 21 to 50 percent category.

The table includes the distributions for workers with narrow and broad definitions of occupational STEM skills. Based on survey responses to whether technical expertise in certain STEM disciplines is required for their job in the 2010 questionnaire, narrow STEM refers to application of knowledge from engineering, computer science, math, and the natural sciences. Broad STEM additionally incorporates knowledge application of skill from social science disciplines. The results show little difference between the distributions for narrow and broad STEM skills. They are both less likely to link to LEHD values more than 20 percent different from the survey value. Therefore, STEM workers appear to have slightly better quality linkages to LEHD than non-STEM workers.

7.3.3 Over- and Under-reporting of Survey Earnings Values

Table 3 examines the relationship between linked earnings values via the distribution of earnings ratio values. Specifically, this information presents the frequency of NSCG under- and over-reporting the LEHD value. Summing across columns, on average, the NSCG has a slight tendency to underreport the LEHD value with 51.59 percent of cases yielding an earnings ratio value less than one. About a quarter of the linked sample has an earnings ratio between 0.75 and 0.95. That result holds across most respondent characteristics covered in this table. Exceptions to that pattern occur among the same groups discussed in the previous table, the youngest and oldest workers, part time workers, and low-income individuals. Respondents with these characteristics most frequently overreport by a large margin with earnings ratios greater than

1.25. The intuition for this behavior carries over from the discussion of the previous table. Low-labor market attachment bears significant influence on the quality of matches among LEHD data.

Table 3: Distribution of Under- and Over-reporting of Annual Earnings Values

		NSCG annual earnings/LEHD annual earnings =					
		< 0.75	[0.75-0.95)	[0.95-1.00)	(1.00-1.05]	(1.05-1.25]	> 1.25
	N	NSCG underreports			NSCG overreports		
Overall	51,000	7.17	25.02	19.40	14.86	14.12	18.12
Sex							
Male	29,000	7.02	24.42	19.42	15.31	14.64	17.93
Female	21,500	7.38	25.83	19.37	14.26	13.43	18.36
Age							
Less than 30	5,800	6.26	20.93	17.68	13.54	15.32	25.28
30-49	25,500	6.51	25.85	19.66	15.58	15.15	16.11
50-64	17,000	7.97	25.93	20.24	14.90	12.46	17.11
65+	2,700	10.42	20.22	15.34	10.57	12.21	28.03
Race/Ethnicity							
White	31,500	6.56	24.87	20.14	15.34	14.16	17.63
Black	4,800	7.18	24.65	19.05	14.29	14.21	19.92
Asian	8,300	9.05	25.82	18.00	14.16	13.38	17.76
NHPI	250	10.22	25.78	17.78	12.89	14.22	18.22
AIAN	200	8.53	28.91	12.32	14.69	11.85	22.75
Multiple race	1,000	6.58	20.84	18.74	15.75	15.15	22.23
Hispanic (of any race)	4,100	7.95	25.83	17.74	13.32	14.93	19.19
Citizenship							
U.S. citizen	47,000	7.10	25.04	19.51	14.88	14.19	18.03
Non-citizen	3,600	8.02	24.76	17.97	14.59	13.27	19.26
Educational Attainment							
Undergraduate degree	27,500	7.33	25.40	19.33	14.61	14.78	17.21
Graduate degree	23,500	6.98	24.57	19.47	15.16	13.36	19.18
Hours Worked*							
Full time	41,500	6.19	26.42	20.88	16.07	14.40	14.96
Part time	5,500	10.09	18.90	12.39	8.66	13.64	34.27
STEM Occupation							
STEM job (narrow)	23,000	6.29	27.45	21.63	16.53	13.90	13.11
Non-STEM job	28,000	7.89	23.04	17.58	13.50	14.31	22.19
STEM job (broad)	29,000	6.11	26.48	21.07	16.23	14.13	14.91
Non-STEM job	22,000	8.56	23.10	17.21	13.07	14.12	22.32
LEHD Earnings Distribution Position							
Quintile 1	10,000	7.94	11.49	7.86	6.00	10.68	53.96

Table 3: Distribution of Under- and Over-reporting of Annual Earnings Values

		NSCG annual earnings/LEHD annual earnings =					
		< 0.75	[0.75-0.95)	[0.95-1.00)	(1.00-1.05]	(1.05-1.25]	> 1.25
	N	NSCG underreports			NSCG overreports		
Quintile 2	10,000	4.86	22.96	20.58	15.58	18.31	16.49
Quintile 3	10,000	4.73	26.88	24.32	18.28	16.34	8.23
Quintile 4	10,000	4.80	29.73	24.95	19.57	13.92	5.96
Quintile 5	10,000	13.53	34.02	19.28	14.88	11.37	5.94

Source: 2010 NSCG and 2011 LEHD EHF files. Results rounded or suppressed (D) where necessary for disclosure avoidance.

Looking at the modal frequencies across the earnings percentiles, this table also supports research findings that survey overreporting occurs at the low end of the earnings distribution while underreporting occurs at the high end[8]. The NSCG reports 25 percent or higher values for over half of the linked cases in the lowest earnings quintile. The frequency of NSCG values underreporting by taking on a value between 0.75 and 0.95 increases monotonically across the remaining quintiles.

7.3.4 Regression Analysis of Measurement Error

Measurement error potentially biases statistical analyses using survey data. For example, Duncan and Hill's (1985) early attempt at measuring bias from measurement error in annual earnings regressions finds error causes estimates to understate the true effect of tenure on earnings by about thirty percent. Bound et al. (2000), highlight the need for more research on bias correction via assessment of correlations between worker characteristics and the measurement error of a variable of interest such as earnings.

The analysis thus far yields insight into which respondent characteristics may be related to the size of measurement error among earnings data. Specifically, measures associated with age, income, and employment volatility should have significant effects. Tables 2 and 3 show noticeable differences between STEM and non-STEM workers. Therefore, we may expect that regression analysis may show significance based on the types of skills used on the job.

Table 4 presents regression results of log-linear models of the earnings ratio (measurement error) as a function of various socioeconomic characteristics of the respondent including employment sector, job tenure and household structure in addition to the characteristics shown above. The first column shows results for the benchmark model that includes all linked cases with non-missing information.

[8] (Bee, 2013; Roemer, 2002; Brummet et al., 2017)

Table 4: Earnings Differential as a Function of Demographic and Employment Characteristics

	Dependent variable: Log(NSCG annual earnings/ LEHD annual earnings)		
	Benchmark	Benchmark (with broad STEM definition)	Marginal workers removed
LEHD earnings quintile 2	-0.759 ***	-0.760 ***	-0.371 ***
	0.031	0.031	0.019
Earnings quintile 3	-0.840 ***	-0.840 ***	-0.437 ***
	0.034	0.034	0.023
Earnings quintile 4	-0.896 ***	-0.895 ***	-0.482 ***
	0.036	0.036	0.023
Earnings quintile 5	-0.996 ***	-0.995 ***	-0.541 ***
	0.037	0.037	0.024
Male	0.086 ***	0.087 ***	0.040 ***
	0.007	0.007	0.003
Age: 30-49	0.045 ***	0.046 ***	0.049 ***
	0.013	0.013	0.007
Age: 50-64	0.045 ***	0.046 ***	0.062 ***
	0.012	0.012	0.008
Age: 65+	0.040 *	0.042 *	0.070 ***
	0.022	0.022	0.018
Black	-0.010	-0.011	-0.024 ***
	0.013	0.013	0.008
Asian	-0.024 ***	-0.022 **	-0.011
	0.009	0.009	0.007
AIAN	-0.103 *	-0.102 *	-0.084 **
	0.052	0.052	0.032
NHPI	0.045	0.045	0.003
	0.042	0.042	0.046
Multiple race	0.014	0.014	0.005
	0.019	0.019	0.009
Hispanic	-0.027 *	-0.027 *	-0.030 ***
	0.015	0.015	0.009
U.S. citizen	0.053 ***	0.052 ***	0.015
	0.010	0.010	0.009
Different employer within 2 years	0.084 ***	0.084 ***	0.117 ***
	0.015	0.015	0.013
Previously retired	-0.036 *	-0.036 *	
	0.020	0.021	
Self-employed	0.176 ***	0.176 ***	
	0.009	0.009	
Private, non-profit	-0.023 **	-0.027 **	-0.031 ***
	0.011	0.011	0.006

Table 4: Earnings Differential as a Function of Demographic and Employment Characteristics

	Dependent variable: Log(NSCG annual earnings/ LEHD annual earnings)		
	Benchmark	Benchmark (with broad STEM definition)	Marginal workers removed
Public	-0.049 ***	-0.053 ***	-0.066 ***
	0.010	0.010	0.008
Other sector of employment	-0.066	-0.067	
	0.097	0.097	
Full time	0.204 ***	0.204 ***	
	0.018	0.018	
Married	0.039 ***	0.039 ***	0.038 ***
	0.010	0.010	0.006
Divorced	0.031 **	0.031 ***	0.035 ***
	0.012	0.011	0.009
Minor child(ren) in household	-0.003	-0.003	0.001
	0.006	0.007	0.004
Master's degree	0.086 ***	0.083 ***	0.048 ***
	0.009	0.008	0.005
Ph.D.	0.136 ***	0.132 ***	0.093 ***
	0.014	0.014	0.011
Professional degree	0.207 ***	0.207 ***	0.147 ***
	0.013	0.013	0.012
STEM skills (narrow)	0.037 ***		0.028 ***
	0.008		0.007
STEM skills (broad)		0.043 ***	
		0.007	
N	37,500	37,500	25,000
R-squared	0.222	0.223	0.187
RMSE	0.567	0.567	0.283

Source: 2010 NSCG and 2011 LEHD EHF files.
**** (p-value ≤ 0.01), ** (p-value ≤ 0.05), *(p-value ≤ 0.1).*
Robust standard errors clustered at the state level. Results rounded or suppressed (D) where necessary for disclosure avoidance.

As expected, the earnings ratio decreases as income increases, and does so at an increasing rate. This result confirms increased survey underreporting among higher income respondents. Age has statistically significant positive effects on the earnings ratio that also reflect the U-shape behavior captured among the youngest and oldest workers in Tables 2 and 3. Workers with short job tenure, or who are self-employed tend to overreport their earnings. Another significant result occurs across educational attainment. Disaggregating graduate level education shows the earnings ratio rises through the attainment of a terminal degree. This result may be associated with the effects of income, as earnings tend to rise with education.

In the benchmark model, STEM as a positive statistically significant effect on the ratio, although it is small and not robust across subsample estimates. Using a broader definition of STEM skills does not change its effect or the significance or magnitude of any other covariates as shown in column 2. Therefore, remaining results will only include the narrowly-defined STEM measure.

The third column of Table 4 shows how the results change when excluding marginal workers from the sample, known to have less coverage within the LEHD framework. This subsample intends to represent individuals with highly stable employment.[9] Most covariates retain their statistical significance from the benchmark model. However, the magnitude of some of those effects change dramatically. For example, the effects from the earnings quintiles and sex are half the size of the benchmark estimates. Aging takes on a monotonically increasing effect on the earnings ratio in the absence of marginal workers. Similarly, the effect of recent employer change is more pronounced. As with earnings, stable employment has a tempering effect on the size of effects from educational attainment.

Additional subsample estimates provide further insight into the behavior of measurement error for selected subgroups: age, part time and full time, and STEM/non-STEM. Formal results for the subsample estimates are in the appendix. As shown above, volatility in measurement error frequently occurs among the youngest and oldest workers. In addition, statistically significant effects among controls describing decreased employment stability such as short job tenure and self-employment increase measurement error in earnings. For example, job tenure less than two years generates its largest effects among workers in their twenties who likely made recent entry into workforce and those aged 65 and older late sixties and seventies either transitioning to full-time or partial retirement.[10] Self-employment retains statistical significance across all age group subsamples, with particularly large effects for older workers. STEM skills are not statistically significant across all age groups, but do yield their largest augmenting effect on very young workers in their twenties. Effect sizes are in general larger for the part time subsample relative to the full time subsample, potentially reflecting volatility in earnings within this group.[11] A similar description applies to the diverse group of non-STEM workers compared to STEM workers.

VIII. CONCLUSION

Summary of Results

This research evaluates conceptual alignment, coverage, and agreement of annual earnings information between the National Survey of College Graduates (NSCG) and Longitudinal Employer-Household Dynamics (LEHD). These datasets were linked by PIK in order to achieve a person-job-level file from which to derive a comparable value for annual earnings across all jobs from the administrative data. The LEHD data provides very good coverage of the NSCG sample (93.95%). However, there are known data limitations, particularly the

[9] Respondents in this subsample are older than 25 years and employed full time. They are not self-employed or working in an undefined sector of the economy. They have never retired, and do not have primary employment in the federal government. Lastly, the dataset is trimmed to include observations within the middle 90% of the earnings distribution.

[10] See Table 5a.

[11] See Table 5b.

underrepresentation of certain groups such as very young workers, retirees, the self-employed and part time workers. This prompted us to carefully consider data performance among the marginally employed throughout the analysis.

Analysis of agreement between linked values found about a third (35.54%) of the linked LEHD data is within five percent of the NSCG value. The youngest and oldest workers, part time workers, and low-income individuals link to much larger LEHD earnings values at greater than average rates. Generally, the NSCG understates the LEHD value, with just over half of the linked sample taking on an NSCG-LEHD earnings ratio value less than one.

Workers at either end of the earnings distribution also displayed notable deviations from average trends in agreement. Specifically low-income respondents have a tendency to overreport their earnings and high-income respondents underreport relative to administrative record values. Distribution results also found STEM workers have better quality matches by more frequently linking to LEHD values closer in value than non-STEM workers do.

Regression analysis found position in the earnings distribution, sex, age, job tenure, self-employment and educational attainment have a statistically significant influence on the earnings ratio (measurement error). These results are robust across several subsample estimates, once again displaying greater volatility within marginal groups.

Recommendations for Future Work

The LEHD is a highly useful data source for earnings information among formally employed individuals of prime working age. Essentially, the more stable the employment, the better the quality of the AR data. The continued quality and usefulness of the LEHD depends on the Census Bureau maintaining relationships with data providers, requiring the Bureau's commitment to implementing privacy protection policies and safeguarding against misuse in exchange for access to timely, quality administrative data on employment.

Data limitations stem from under-coverage of certain occupations and groups of people associated with marginal employment and/or informal payment practices. Integration into the formal labor market is a highly significant factor in the determination of quality linkages to AR data that may be used for survey supplementation.

Future work should focus on overcoming the aforementioned data limitations. For example, investigating alternative data sources on earnings that provide similar coverage and quality such as federal tax data from IRS forms W-2 and 1040 has the added benefit of requiring only one agreement to gain access to nationwide employment information. Given a well-established relationship exists between the Census Bureau and the Internal Review Service, regular access to this information should not be an issue. The results of its analysis would also serve as a robustness check to what was observed in this study using LEHD information. Tax data may also circumvent the omission of certain federal employees from the universe of available data. While LEHD documentation cites heightened sensitivity among employees in defense and judicial agencies, this privilege likely does not extend to the neglect of proper documentation for taxation. Finally, tax data provides information on household income for spouses filing jointly

allowing even more insight into additional sources of income from shared assets such as a business.

Another suggestion for future work involves conducting a sensitivity analysis on the earnings distribution to observe how measures of dispersion change when replacing survey values with AR values. This research would begin by recording relevant descriptive statistics for the NSCG earnings distribution, and then derive those same statistics for a modified distribution replacing all or some NSCG values with AR values for comparison. By isolating value replacement by respondent characteristics the analysis allows rigorous analysis of earnings data for marginal workers, for example, and sheds light on the effects of undercoverage for cases where LEHD did not verify employment and would replace survey data with missing values.

IX. REFERENCES

Abraham, Katharine G., John Haltiwanger, Kristin Sandusky, James R. Spletzer. 2013. "Exploring Differences in Employment between Household and Establishment Data". *Journal of Labor Economics* 31, no. 2: 129-s172.

Abowd, John M., Bryce E. Stephens, Lars Vilhuber, Fredrik Andersson, Kevin L. McKinney, Marc Roemer, and Simon Woodcock. 2009. "The LEHD Infrastructure Files and the Creation of the Quarterly Work-Force Indicators". In *Producer Dynamics*, ed. Timothy Dunne, J. Brad-ford Jensen, and Mark J. Roberts. Chicago: University of Chicago Press.

Bee, Adam. 2013. "An Evaluation of Retirement Income in the CPS ASEC Using Form 1099-R Microdata". Working Paper. U.S. Census Bureau, Washington, D.C. https://www.census.gov/library/working-papers/2013/demo/Bee-PAA-paper.html

Bollinger, Christopher R. and Martin H. David. 2001. "Estimation with Response Error and Nonresponse: Food-Stamp Participation in the SIPP". *Journal of Business and Economic Statistics*. Vol. 19 no. 2. 129-141.

Bollinger, Christopher R., Barry T. Hirsch, Charles M. Hokayem, and James P. Ziliak. 2015. "Trouble in the Tails? Earnings Nonresponse and Response Bias across the Distribution". Working paper. http://economics.emory.edu/home/documents/Seminars%20Workshops/Seminar-2015-Hirsch.pdf

Bound, John and Alan B. Krueger. 1991. "The Extent of Measurement Error in Longitudinal Earnings Data: Do Two Wrongs Make a Right?" *Journal of Labor Economics*. Vol. 9 no. 1. 1-24.

Bowie, Chester and Daniel Kasprzyk. 1987. "A Review of Administrative Records in the Survey of Income and Program Participation". SEHSD Working Paper no. 8721-43, U.S. Census Bureau, Washington, D.C.

Brummet, Quentin, Denise Flanagan-Doyle, Joshua Mitchell, John Voorheis, Laura Erhard, and Brett McBride. 2018. "Investigating the Use of Administrative Records in the Consumer Expenditure Survey". CARRA Working Paper. US Census Bureau, Washington, DC.

Dillon, Michaela. 20XX. "Evaluating Administrative Records to Inform Measurement Error Properties of National Survey of College Graduates Estimates: Employment History and Firm Characteristics". CARRA Technical Memorandum. U.S. Census Bureau, Washington, D.C.

Duncan, Greg J. and Daniel H. Hill. 1985. "An Investigation of the Extent and Consequences of Measurement Error in Labor-Economic Survey Data". *Journal of Labor Economics*. Vol. 3 no. 4. 508-532.

Korinek, Anton, Johan A. Mistiaen, and Marin Ravallion. 2005. "Survey Nonresponse and the Distribution of Income". World Bank Policy Research Working Paper no. 3543. World Bank, Washington, D.C.

Künn, Steffen. 2015. "The Challenges of Linking Survey and Administrative Data". IZA World of Labor. https://wol.iza.org/articles/challenges-of-linking-survey-and-administrative-data/long

Kreuter, Frauke, Gerrit Muller, and Mark Trappmann. 2014. "A Note on Mechanisms Leading to Lower Data Quality of Late or Reluctant Respondents". *Sociological Methods and Research*. Vol. 43 no. 3. 452-464.

Lillard, Lee, James P. Smith, and Finis Welch. 1986. “What Do We Really Know about Wages? The Importance of Nonreporting and Census Imputation”. *Journal of Political Economy*. Vol. 94 no. 3. 489-506.

Meyer, Bruce D., Wallace K. C. Mok, and James X. Sullivan. 2015. “Household Surveys in Crisis”. *Journal of Economic Perspectives*. vol. 29 no. 4: 199-226.

Roemer, Mark. 2002. “Using Administrative Earnings Records to Assess Wage Data Quality in the March Current Population Survey and the Survey of Income and Program Participation”. Working Paper. U.S. Census Bureau, Washington, D.C. https://www.census.gov/content/dam/Census/library/working-papers/2002/demo/asa2002.pdf

Vilhuber, Lars and Devin McKinney. 2014. “LEHD Infrastructure Files in the Census RDC—Overivew”. CES Working Paper no. 14-26. US Census Bureau, Washington, DC.

Wagner, Deborah and Mary Layne. 2014. “The Person Identification Validation System (PVS): Applying the Center for Administrative Records Research and Applications’ (CARRA) Record Linkage Software”. CARRA Working Paper no. 2014-01. U.S. Census Bureau, Washington, D.C.

X. APPENDIX

Figure 3: Derivation of LEHD Annual Earnings Value

Panel A: Full linked dataset

NSCG		LEHD							
EARN ($1,000)	PIK	Year	SEIN	SEINUNIT	Q1	Q2	Q3	Q4	Annual
80	A	2009	001	2	17.5	17.5	17.5	17.5	70
80	A	2009	002	1	-	-	5	5	10
80	A	2010	001	2	17.85	17.85	17.85	17.85	71.4
95	B	2008	123	1	23.75	23.75	23.75	23.75	95
95	B	2009	123	1	23.75	23.75	23.75	24.25	95.5

Panel B: Remove records for years other than 2009

NSCG		LEHD							
EARN ($1,000)	PIK	Year	SEIN	SEINUNIT	Q1	Q2	Q3	Q4	Annual
80	A	2009	001	2	17.5	17.5	17.5	17.5	70
80	A	2009	002	1	-	-	5	5	10
95	B	2009	123	1	23.75	23.75	23.75	24.25	95.5

Panel C: Sum annual earnings by PIK

NSCG		LEHD								
EARN ($1,000)	PIK	Year	SEIN	SEINUNIT	Q1	Q2	Q3	Q4	Annual	Earnings Sum
80	A	2009	001	2	17.5	17.5	17.5	17.5	70	70
80	A	2009	002	1	-	-	5	5	10	80
95	B	2009	123	1	23.75	23.75	23.75	24.25	95.5	95.5

Panel D: Retain last obs. for each PIK

NSCG		LEHD								
EARN ($1,000)	PIK	Year	SEIN	SEINUNIT	Q1	Q2	Q3	Q4	Annual	Earnings Sum
80	A	2009	002	1	-	-	5	5	10	80
95	B	2009	123	1	23.75	23.75	23.75	24.25	95.5	95.5

Panel E: Reduced dataset for earnings differential analysis

NSCG		LEHD								
EARN ($1,000)	PIK	Year	SEIN	SEINUNIT	Q1	Q2	Q3	Q4	Annual	Earnings Sum
80	A	2009	002	1	-	-	5	5	10	80
95	B	2009	123	1	23.75	23.75	23.75	24.25	95.5	95.5

Source: 2010 NSCG and 2011 LEHD EHF files.

- Panel A depicts an example snapshot of the linked NSCG-LEHD file. The columns EARN and PIK come from the NSCG. The LEHD provides PIK and the information in the remaining columns. Note there are multiple observations per PIK, as the LEHD provides information for each job held by the individual over time.
- In Panel B, since NSCG requests 2009 annual earnings, all observations associated with years other than 2009 are deleted.
- Currently, each value of EARN is an aggregated value of several sources of income linked to an annual earnings value from one single firm. In order to equate the NSCG and LEHD annual earnings values in the sense of an aggregate value, the LEHD "Annual" values for each PIK must be summed. Panel C shows respondent with PIK value A reported $80,000 in earnings from all sources of employment wages to the NSCG. The LEHD shows two sources of earnings for respondent A in 2009, Firm (SEIN) 001 and 002 paid the respondent $70,000 and $10,000, respectively. A new variable, "Earnings Sum" is created to keep a running total of LEHD "Annual" values for each PIK, such that the last observation for respondent A is the sum of his earnings from Firms 001 and 002, $80,000.
- Once the summed value of annual earnings is calculated for each PIK, retain the final observation for each PIK. This is shown in Panel D with the deletion of the first observation for respondent A.
- Panel E shows the reduced dataset retaining one observation for each PIK reporting the aggregated annual earnings value for both the NSCG and LEHD datasets. The analysis compares the values of NSCG's EARN and the LEHD-derived "Earnings Sum". Assuming LEHD data is the "true" value, there is no measurement error for respondent A as both data sources report $80,000. There is error, however, for respondent B who reported $95,000 to NSCG while LEHD reports $95,500.

Table 5a: Subsample Estimates of the Earnings Differential as a Function of Demographic and Employment Characteristics

	Dependent variable: Log(NSCG annual earnings/LEHD annual earnings)									
	Benchmark		Age: Less than 30		Age: 30-49		Age: 50-64		Age: 65+	
LEHD earnings quintile 2	0.759	***	0.683	***	0.822	***	0.744	***	0.545	***
	0.031		0.044		0.042		0.041		0.058	
Quintile 3	0.840	***	0.756	***	0.901	***	0.832	***	0.654	***
	0.034		0.052		0.044		0.043		0.057	
Quintile 4	0.896	***	0.806	***	0.955	***	0.889	***	0.708	***
	0.036		0.069		0.045		0.045		0.060	
Quintile 5	0.996	***	0.848	***	1.045	***	0.998	***	0.876	***
	0.037		0.128		0.046		0.044		0.057	
Male	0.086	***	0.080	***	0.080	***	0.087	***	0.135	***
	0.007		0.026		0.008		0.014		0.046	
Age: 30-49	0.045	***								
	0.013									
Age: 50-64	0.045	***								
	0.012									
Age: 65+	0.040	*								
	0.022									
Black	0.010		0.019		0.004		0.034		0.090	
	0.013		0.038		0.015		0.023		0.081	
Asian	0.024	***	0.070	**	0.010		0.076	***	0.055	
	0.009		0.028		0.011		0.023		0.033	
AIAN	0.103	*	(D)		0.112		0.071		(D)	
	0.052				0.078		0.102			
NHPI	0.045		0.232		0.020		0.024		(D)	
	0.042		0.246		0.053		0.112			
Multiple race	0.014		0.065		0.011		0.004		0.086	
	0.019		0.042		0.030		0.037		0.096	
Hispanic	0.027	*	0.035		0.015		0.038	*	0.008	
	0.015		0.031		0.020		0.020		0.046	
U.S. citizen	0.053	***	0.097	***	0.058	***	0.053	***	0.100	
	0.010		0.032		0.013		0.019		0.059	
Different employer within 2 years	0.084	***	0.113	***	0.067	***	0.090	***	0.167	*
	0.015		0.036		0.018		0.021		0.086	
Previously retired	0.036	*	(D)		0.079	*	0.000		0.046	
	0.020				0.041		0.029		0.040	
Self-employed	0.176	***	0.073	*	0.147	***	0.202	***	0.353	***
	0.009		0.038		0.018		0.019		0.037	
Private, non-profit	0.023	**	0.037		0.035	**	0.016		0.106	**
	0.011		0.037		0.015		0.019		0.051	

Table 5a: Subsample Estimates of the Earnings Differential as a Function of Demographic and Employment Characteristics

	Dependent variable: Log(NSCG annual earnings/LEHD annual earnings)									
	Benchmark		Age: Less than 30		Age: 30-49		Age: 50-64		Age: 65+	
Public	0.049	***	0.039		0.067	***	0.042	***	0.047	
	0.010		0.033		0.012		0.013		0.039	
Other sector of employment	0.066		(D)		0.166	**	(D)		(D)	
	0.097				0.074					
Full time	0.204	***	0.033		0.222	***	0.231	***	0.196	***
	0.018		0.062		0.022		0.031		0.049	
Married	0.039	***	0.036	**	0.047	***	0.041	**	0.011	
	0.010		0.015		0.013		0.021		0.072	
Divorced	0.031	**	0.092		0.025		0.039	**	0.013	
	0.012		0.173		0.018		0.018		0.076	
Minor child(ren) in household	0.003		0.045	**	0.002		0.003		0.084	
	0.006		0.019		0.008		0.011		0.099	
Master's degree	0.086	***	0.139	***	0.079	***	0.078	***	0.115	***
	0.009		0.023		0.011		0.012		0.040	
Ph.D.	0.136	***	0.164		0.129	***	0.147	***	0.141	**
	0.014		0.130		0.018		0.022		0.057	
Professional degree	0.207	***	0.136	**	0.176	***	0.222	***	0.331	***
	0.013		0.064		0.016		0.023		0.068	
STEM skills (narrow)	0.037	***	0.129	***	0.024	***	0.028	**	0.043	
	0.008		0.037		0.007		0.011		0.027	
N	37,500		3,600		19,500		13,000		1,700	
R-squared	0.222		0.217		0.236		0.217		0.208	
RMSE	0.567		0.653		0.533		0.578		0.649	

Source: 2010 NSCG and 2011 LEHD EHF files.

**** (p-value ≤ 0.01), ** (p-value ≤ 0.05), *(p-value ≤ 0.1).*

Robust standard errors clustered at the state level. Results rounded or suppressed (D) where necessary for disclosure avoidance.

Table 5b: Subsample Estimates of the Earnings Differential as a Function of Demographic and Employment Characteristics (continued)

	Dependent variable: Log(NSCG annual earnings/LEHD annual earnings)											
	Benchmark		Full time		Part time		STEM occupation (narrow definition)		Non-STEM occupation (narrow definition)			
LEHD earnings quintile 2	0.759	***	0.880	***	0.439	***	0.978	***	0.643	***		
	0.031		0.044		0.030		0.050		0.030			
Quintile 3	0.840	***	0.956	***	0.478	***	1.066	***	0.719	***		
	0.034		0.048		0.029		0.053		0.032			
Quintile 4	0.896	***	1.006	***	0.576	***	1.122	***	0.760	***		
	0.036		0.049		0.054		0.053		0.036			
Quintile 5	0.996	***	1.103	***	0.702	***	1.219	***	0.863	***		
	0.037		0.050		0.043		0.053		0.037			
Male	0.086	***	0.075	***	0.177	***	0.068	***	0.105	***		
	0.007		0.006		0.030		0.009		0.009			
Age: 30-49	0.045	***	0.067	***	0.115	**	0.040	***	0.060	**		
	0.013		0.013		0.053		0.014		0.028			
Age: 50-64	0.045	***	0.065	***	0.144	**	0.046	**	0.053	**		
	0.012		0.012		0.058		0.018		0.025			
Age: 65+	0.040	*	0.076	***	0.174	**	0.045	**	0.045			
	0.022		0.024		0.068		0.022		0.034			
Black	0.010		0.023	**	0.050		0.011		0.020			
	0.013		0.012		0.067		0.014		0.017			
Asian	0.024	***	0.020	**	0.058	*	0.015		0.018			
	0.009		0.009		0.030		0.011		0.013			
AIAN	0.103	*	0.136	**	(D)		0.241	***	0.010			
	0.052		0.054				0.061		0.069			
NHPI	0.045		0.038		(D)		0.003		0.081			
	0.042		0.052				0.060		0.093			
Multiple race	0.014		0.016		0.002		0.042	*	0.005			
	0.019		0.018		0.074		0.021		0.031			
Hispanic	0.027	*	0.033	**	0.015		0.032	*	0.021			
	0.015		0.015		0.042		0.016		0.019			
U.S. citizen	0.053	***	0.058	***	0.025		0.058	***	0.046			
	0.010		0.009		0.065		0.015		0.034			
Different employer within 2 years	0.084	***	0.047	***	0.167	***	0.046	***	0.112	***		
	0.015		0.015		0.049		0.015		0.024			
Previously retired	0.036	*	0.012		0.036		0.004		0.055	*		
	0.020		0.017		0.044		0.033		0.028			
Self-employed	0.176	***	0.130	***	0.420	***	0.101	***	0.261	***		
	0.009		0.011		0.041		0.015		0.019			

Table 5b: Subsample Estimates of the Earnings Differential as a Function of Demographic and Employment Characteristics (continued)

	Dependent variable: Log(NSCG annual earnings/LEHD annual earnings)											
	Benchmark		Full time		Part time		STEM occupation (narrow definition)		Non-STEM occupation (narrow definition)	
Private, non-profit	0.023	**	0.035	***	0.052		0.075	***	0.009	
	0.011		0.011		0.033		0.013		0.016	
Public	0.049	***	0.045	***	0.052	*	0.085	***	0.025	
	0.010		0.010		0.030		0.010		0.015	
Other sector of employment	0.066		0.035		(D)		0.086		0.025	
	0.097		0.100				0.078		0.137	
Full time	0.204	***					0.252	***	0.164	***
	0.018						0.032		0.022	
Married	0.039	***	0.044	***	0.009		0.028	**	0.049	***
	0.010		0.010		0.040		0.011		0.015	
Divorced	0.031	**	0.045	***	0.087	**	0.017		0.045	***
	0.012		0.013		0.041		0.021		0.015	
Minor child(ren) in household	0.003		0.002		0.000		0.001		0.003	
	0.006		0.006		0.035		0.007		0.010	
Master's degree	0.086	***	0.081	***	0.109	***	0.048	***	0.124	***
	0.009		0.008		0.039		0.011		0.011	
Ph.D.	0.136	***	0.122	***	0.215	**	0.102	***	0.186	***
	0.014		0.013		0.084		0.019		0.028	
Professional degree	0.207	***	0.199	***	0.221	***	0.220	***	0.192	***
	0.013		0.014		0.043		0.027		0.016	
STEM skills (narrow)	0.037	***	0.040	***	0.022					
	0.008		0.006		0.034					
N	37,500		33,500		3,800		20,500		17,000	
R-squared	0.222		0.246		0.151		0.257		0.205	
RMSE	0.567		0.530		0.796		0.514		0.620	

Source: 2010 NSCG and 2011 LEHD EHF files.

**** (p-value ≤ 0.01), ** (p-value ≤ 0.05), *(p-value ≤ 0.1).*

Robust standard errors clustered at the state level. Results rounded or suppressed (D) where necessary for disclosure avoidance.